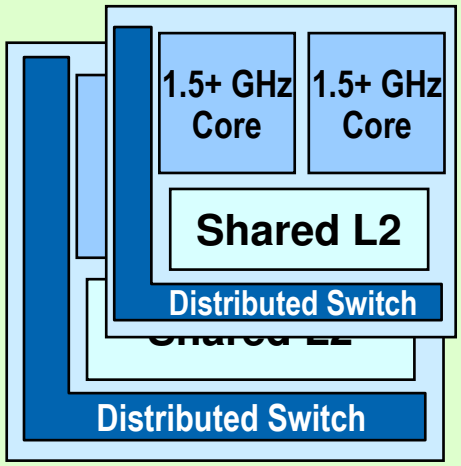
Hardware & Software Roadmap

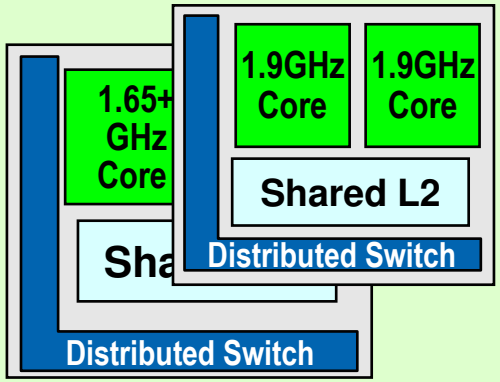
HPC Development

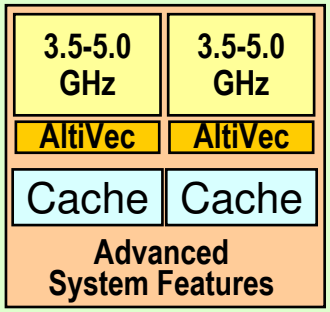
Piyush Chaudhary
piyushc@us.ibm.com

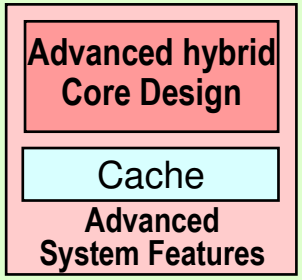
POWER Processor Roadmap

2001-4	2004-6	2007-9	2010-11
POWER4 / 4+	POWER5 / 5+	POWER6	POWER7









Workload Accelerators
Highly threaded cores

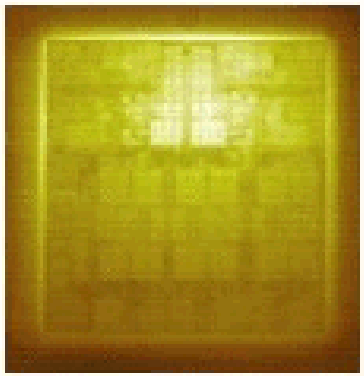
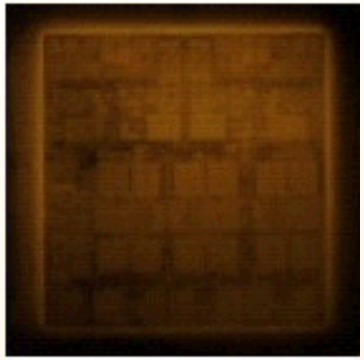
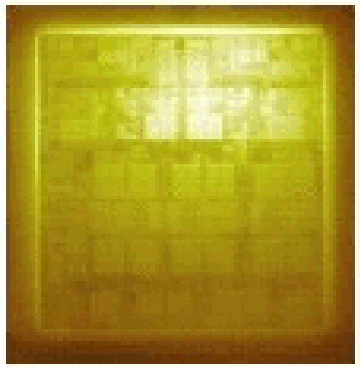
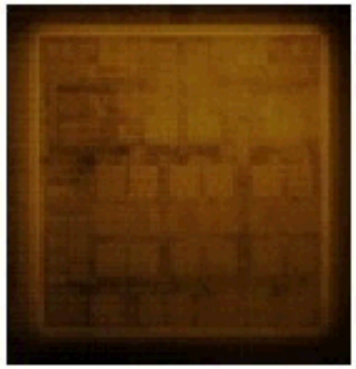
Chip Multi Processing
- Distributed Switch
- Shared L2
Dynamic LPARs (32)

2.3 GHz POWER5+
Enhanced Scaling
Simultaneous Multi-Threading (SMT)
Enhanced Distributed Switch
Enhanced Core Parallelism
Improved FP Performance
Increased memory bandwidth
Reduced memory latencies
Virtualization

Very High Frequencies 4-5GHz
Enhanced Virtualization
Advanced Memory Subsystem
Altivec / SIMD instructions
Instruction Retry
Dynamic Energy Management
Partition Mobility
Memory Protection Keys

BINARY COMPATIBILITY

Dynamic Power Management

	No Power Management	Dynamic Power Management
Single Thread		
Simultaneous Multithreading		

Photos taken with thermal sensitive camera while prototype POWER5 chip was undergoing tests

Simultaneous Multithreading with dynamic power management reduces power consumption below standard, single threaded level

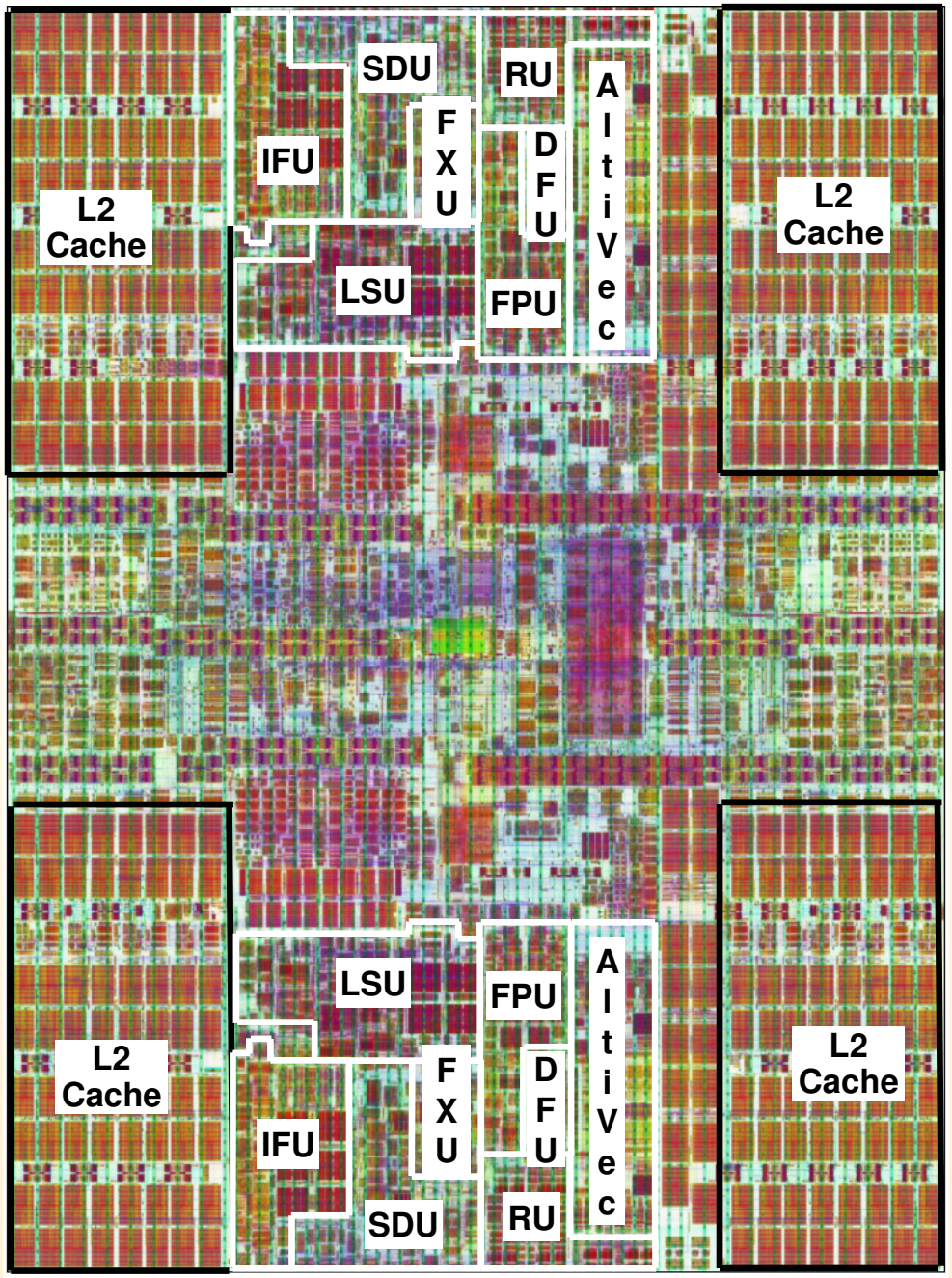
POWER6

POWER6 Chip Overview

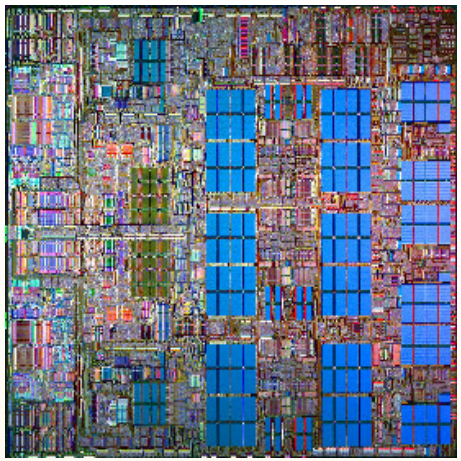
Dual Core chip
- 2-way SMT core
- Ultra-high frequency
  - 3.5 - 5 GHz
- >750M transistors
- Superscalar
- Large on-chip L2
- On-chip L3 directory & controller
- Two memory controllers on-chip

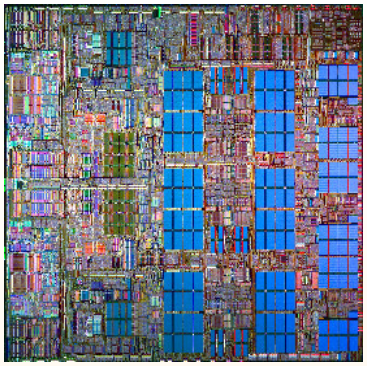
Technology
- CMOS 65nm lithography, SOI

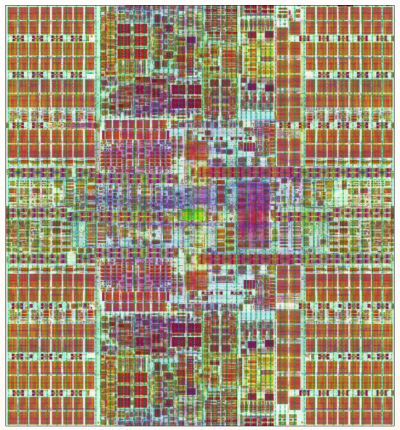
Error checking and recovery



POWER6 Design







	POWER5	POWER5+	POWER6
Size	389 mm²	245 mm²	341 mm²
Transistors	276 M	276 M	>750 M
Cores	2	2	2
Frequencies	1.65 GHz	1.9 GHz	4 GHz
L2 Cache	1.9MB Shared	1.9MB Shared	4MB / Core
L3 Cache	36MB	36MB	32MB
Memory Cntrl	1	1	2 / 1
LPAR	10 / Core	10 / Core	10 / Core

POWER5+ and POWER6 Hierarchy

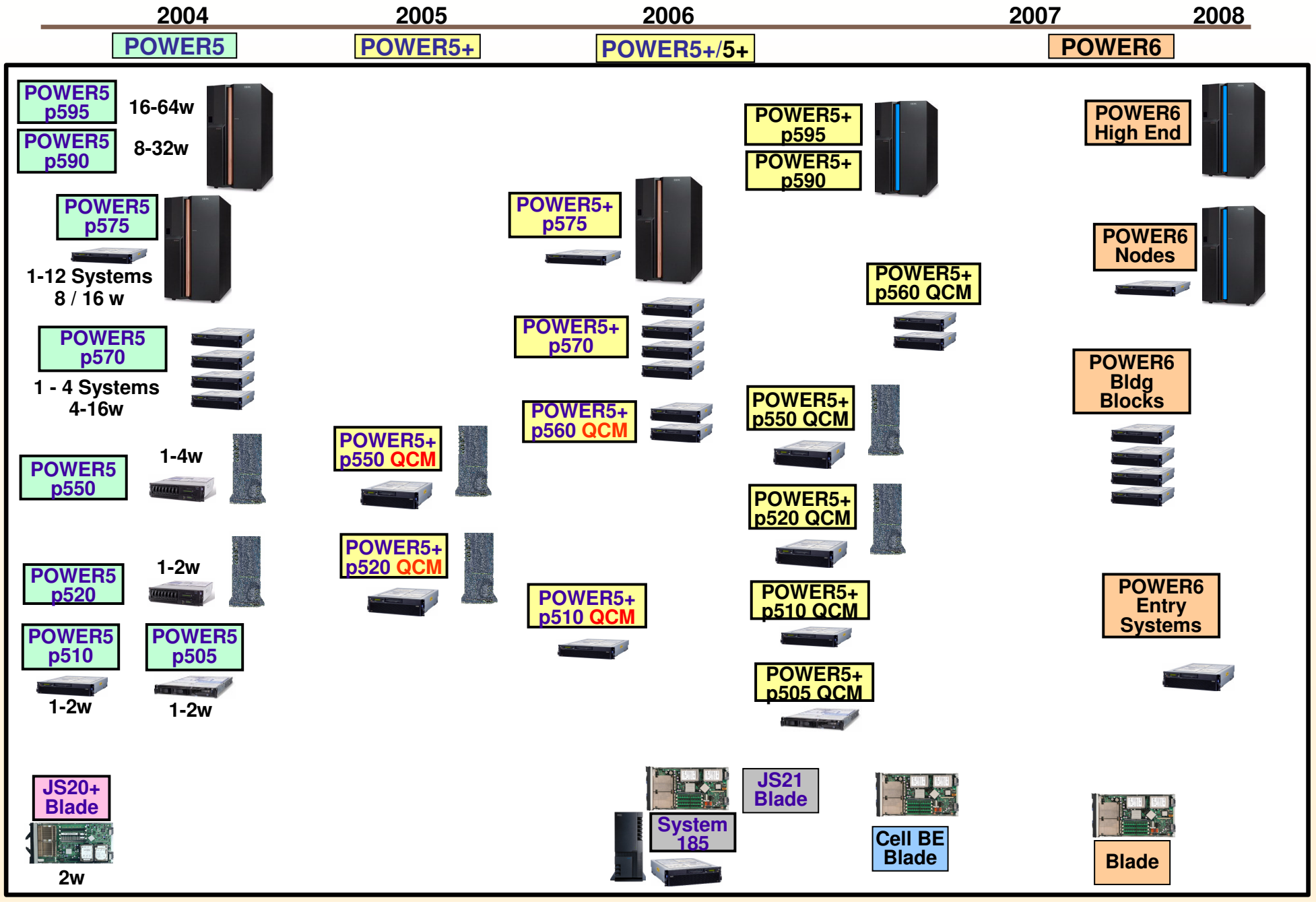
	POWER5+	POWER6
L1 Cache		
ICache capacity, associativity	64 KB, 2-way	64 KB, 4-way
DCache capacity, associativity	32 KB, 4-way	**64 KB, 8-way**
L2 Cache		
Capacity, line size	1.9 MB, 128 B line	**2 x 4 MB, 128 B line**
Associativity, replacement	10-way, LRU	8-way, LRU
Off-chip L3 Cache		
Capacity, line size	36 MB, 256 B line	32 MB, 128 B line
Associativity, replacement	12-way, LRU	**16-way, LRU**
Memory	2 TB maximum	4 TB maximum
Memory bus	2x DRAM frequency	**4x DRAM frequency**

Processor Design

	POWER5+	POWER6
Style	General out-of-order execution	Mostly in-order with special case out-of-order execution
Units	2FX, 2LS, 2FP, 1BR, 1CR	2FX, 2LS, 2FP, 1BR/CR, **1VMX**
Threading	2 SMT threads Alternate ifetch Alternate dispatch (up to 5 instructions)	2 SMT threads Priority-based dispatch **Simultaneous dispatch from two threads (up to 7 instructions)**

Roadmap

Planned IBM System p™ Roadmap

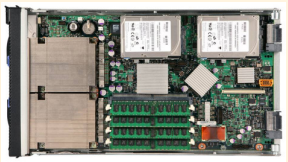


POWER6

Planned POWER6 Systems

Entry System
Up to 8 Core SMP
Rack Server

Blade
2 / 4 Core



Mid Range
Up to 16 Core SMP
4U Rack Optimized
Building Blocks

Compute / Cluster
16 / 32 Core SMP
2U Building Blocks

High End
16 – 64 Core SMP

POWER6 High Lights

Processors

- ~100% higher frequencies
- Enhanced Micro partitioning support
  - Partition Mobility
  - Partition Hibernation
  - Virtual Memory
  - Pooling of processors

New IO

- SAS
- SATA
- SFF
- New IO Drawers

PCIe

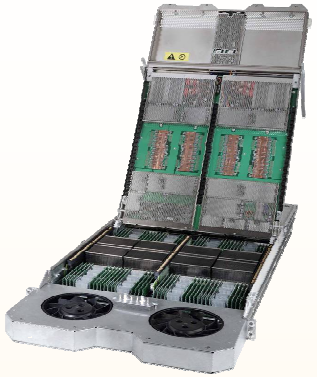
- x8 & x16 implementation
- SAS , Ethernet, & Fiber Channel

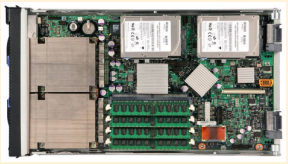
DDR2 memory

- Buffered memory
- Higher memory frequency

Enhanced Planar options

- Faster GX buses
- PCIe, SAS, etc.





POWER6 Virtualization capabilities

Virtual Ethernet Support
- Virtualization of Integrated Ethernet

Partition Mobility
- Partition Availability
- Move “Live” partitions from one physical system to another
- All POWER6 systems

Micro Partition Pooling
- Pool groups of micro partitions for processor resources

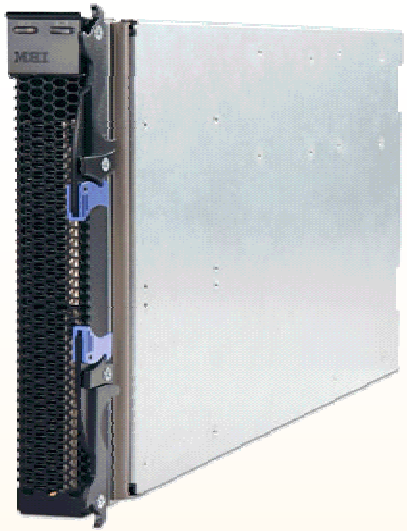
Workload Partitions
- Partitions within partitions
- Single AIX image manages multiple partition images
- Mobility support

Virtual Memory support
- Shared memory pages
- Borrow memory from other partition

Shared Dedicated Processor
- Sharing dedicated processor with the Shared Processor Pool

Partition Hibernation
- Suspend a partition and restart it at a later time

Planned POWER6 Blade



5.3 & 6.1

Linux

POWER6 Blade	
Architecture	Up to 4 Cores
DDR2 Memory	Up to 32GB
DASD / Bays	1 DASD
Expansion	Blade Center I/O
Integrated SAS	Yes
Integrated USB	Yes
Integrated Ports	Blade Center
Integrated Ethernet	BladeCenter
Remote IO Drawers	N/A

Partition Mobility: Active and Inactive LPARs

Active Partition Mobility

- Active Partition Migration is the actual movement of a running LPAR from one physical machine to another without disrupting* the operation of the OS and applications running in that LPAR.

- Applicability
  - Workload consolidation (e.g. many to one)
  - Workload balancing (e.g. move to larger system)
  - Planned CEC outages for maintenance/upgrades
  - Impending CEC outages (e.g. hardware warning received)

Inactive Partition Mobility

- Inactive Partition Migration transfers a partition that is logically 'powered off' (not running) from one system to another.

Partition Mobility supported on POWER6™
AIX 5.3, AIX 5.4 and Linux

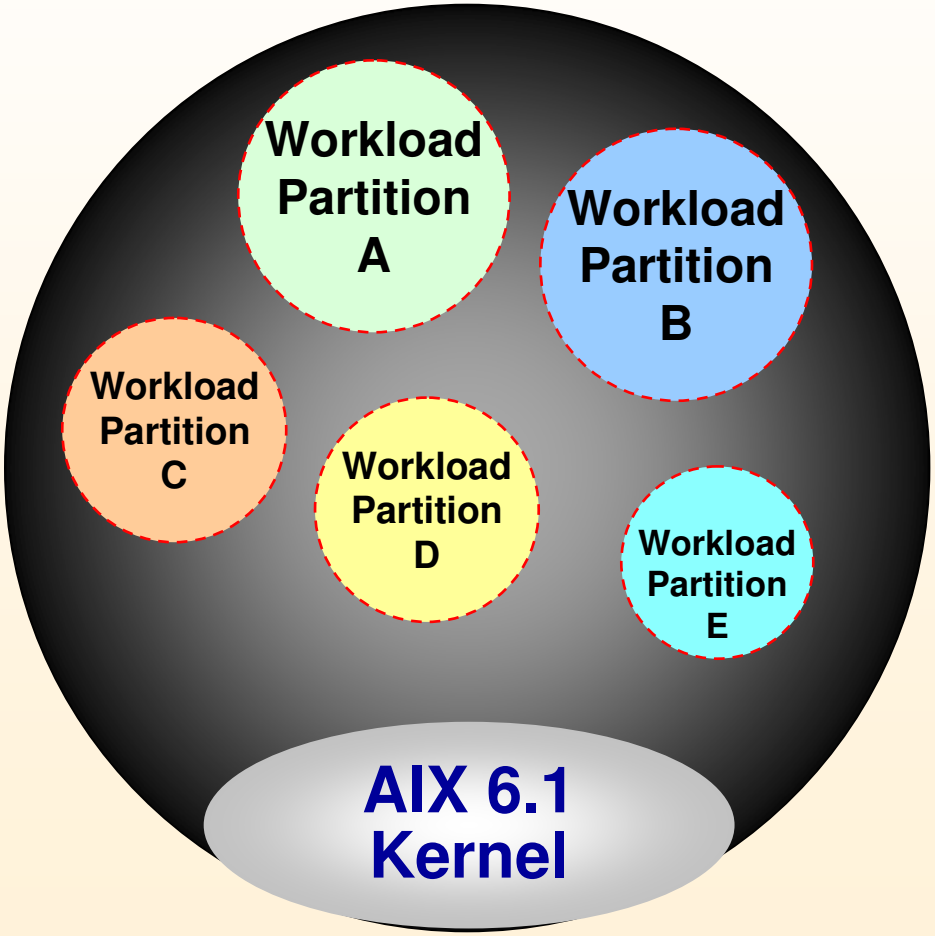
Workload Partitions…

Separate regions of application space within a single AIX image

Partitioned system capacity
- Each Workload Partition obtains a regulated share of the processor and memory resources
- Each Workload Partition has separate network and filesystems and many system services (e.g. telnetd, etc.)

Separate Administrative control
- Each partition is a separate administrative and security domain

Shared system resources
- I/O Devices
- Processor
- Operating system
- Shared Library and Text



I/O

Interconnect Technologies….

Technology	Performance	Processor Environment	Function
RIO	500 MB/sec	RS64 POWER4	Remote I/O support
RIO-2	1 GB/sec	POWER4 POWER5 POWER5+	Remote I/O support
InfiniBand	10 Gb/sec 20 Gb/sec	**POWER5/5+** **POWER6**	Remote I/O support System Interconnect
10 Gb Ethernet	10 Gb/sec	POWER5 POWER5+ **With PCI-X DDR**	System Interconnect
HPS	2 GB/sec	POWER4 POWER5/5+	System Interconnect

P6 & Infiniband Solution Details

- Power6 p575 and Galaxy2 4x DDR solution
  - 04/2008 Planned GA
  - AIX5.3 and SLES10
  - Silverstorm Switch Fabric

- Power6 JS24 Blades and Mellanox PCI-E 4x DDR solution
- 1H/2008 Planned GA
- AIX5.3
- Silverstorm Switch Fabric

- Considering Voltaire Switch Fabric support at a later time

- Planned AIX6.1 upgrade in 10/2008

AIX

AIX 5L Features Roadmap

AIX 5L V5.3– TL4

- Multiple page size
- Support: 4K, 64K, 16M, & 16GB

AIX 5L V5.3– TL5

- LPARs: Common OS Image
- AIX Login (Using LDAP Client)
  - Microsoft Active Directory
- CAPP/EAL4+ Security

AIX 5L V5.3– TL6

- POWER6 Enablement
- Decimal Floating Point

11/2007

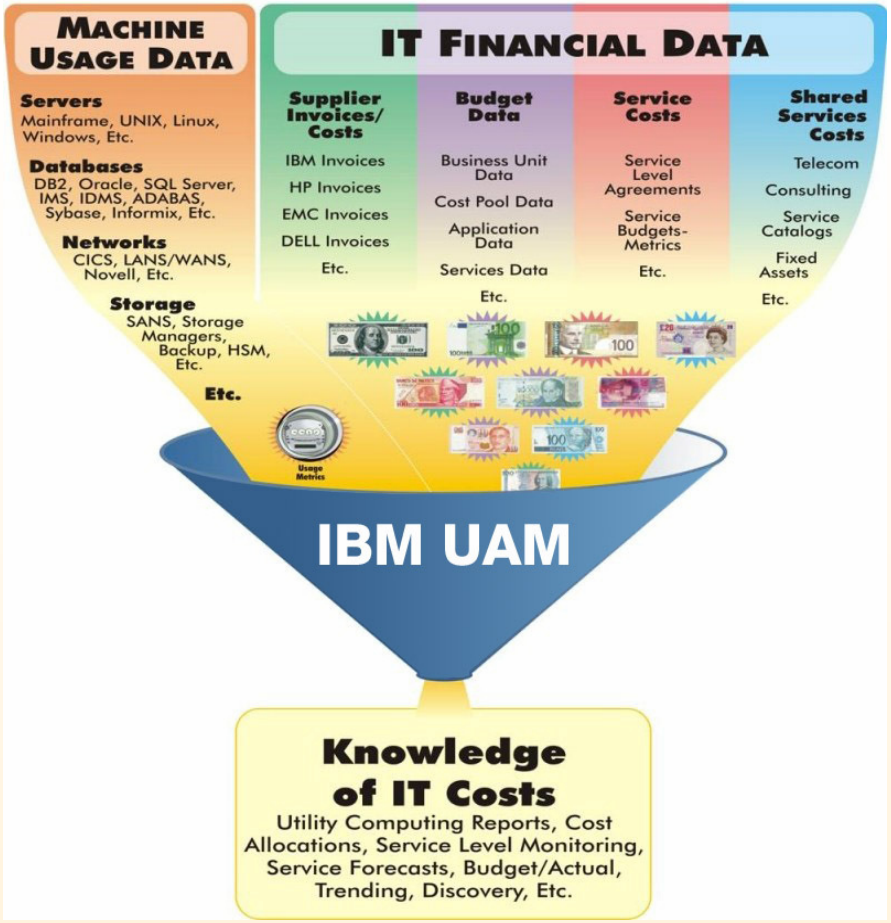
AIX 5L V6.1

- POWER 6 systems & I/O
- Reliability / Availability / Serviceability
  - Storage Keys (Kernel)
  - Concurrent OS Maintenance
  - Enhanced Error Detection / Recovery
- Workload Partitions
  - Mobility support
- Encrypted File System
- LPARs
  - Partition Mobility
  - Partition Hibernation
  - Virtual IO: MPIO Physical & Virtual

All statements regarding IBM future directions and intent are subject to change or withdrawal without notice

IBM Usage & Accounting Manager

- **Helps businesses to understand the true costs of their IT**
  - **Who is consuming which resources?**
  - **What are the true costs of these resources?**
  - **How should costs be allocated for ROI or chargeback?**
- **Enables businesses to make informed decisions about IT options and acquisitions**
- **Facilitates chargeback accounting to bill internal or external customers for their actual resource use**
- **Tracks and analyzes resource utilization across the entire enterprise**
  - **Servers, storage, networks, applications, etc.**



GPFS 3.2

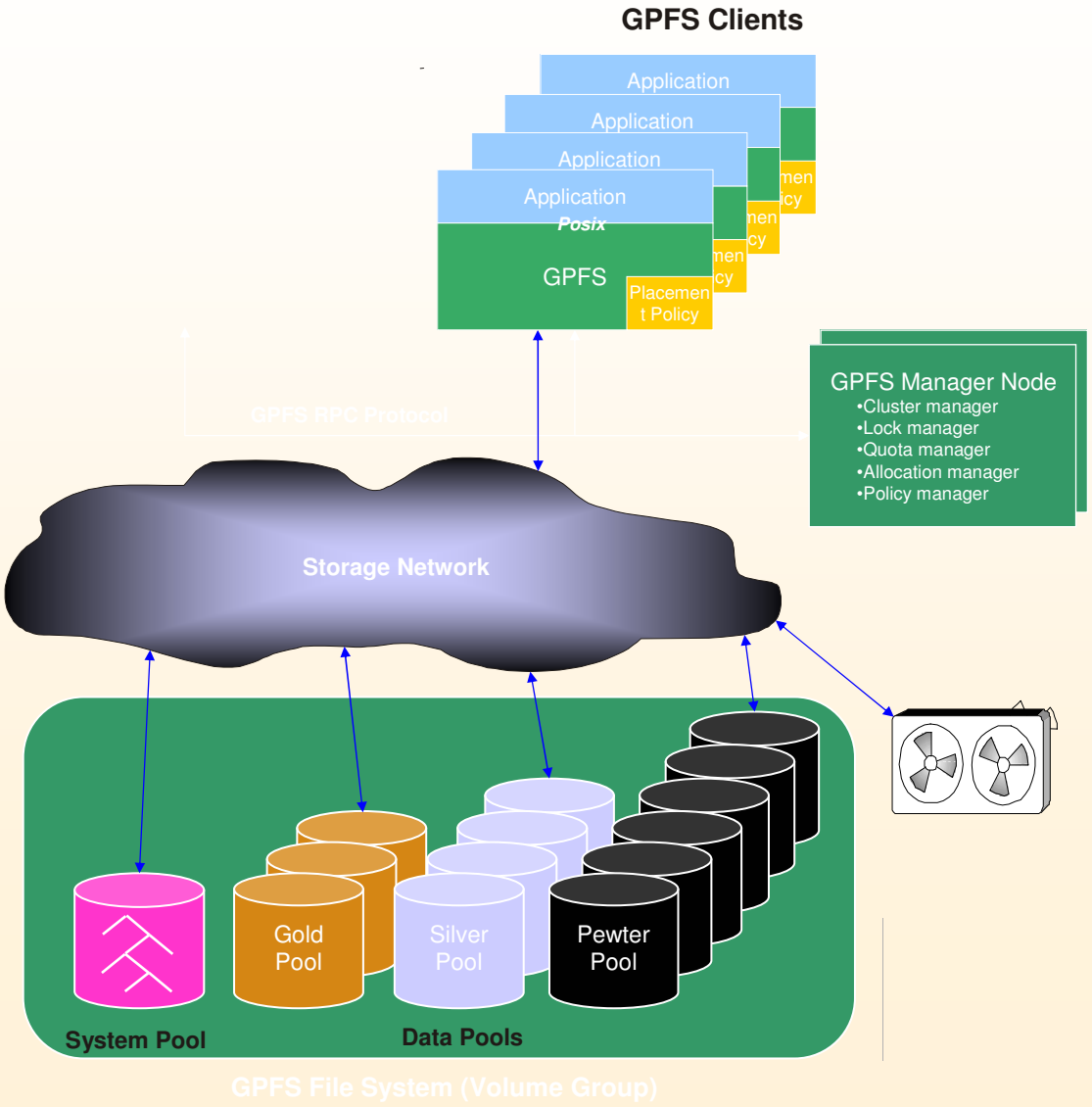
- Policy managed disk-tape migration
- More platforms (Windows, Solaris)
- Infiniband – uDAPL RDMA data transfers
- Admin/Manageability improvements
  - Tracing improvements (mmtracectl command)
  - SNMP support
  - Improved failover times for database usage (persistent reserve)
  - Reconnecting broken sockets
- NSD improvements
  - Multiple active NSD servers per LUN
- Support for NFS4 on Linux
- Scaling and performance improvements
  - Multi-node file create in the same directory
  - Large pagepool support
  - More mounted filesystems (256)
  - SMP scaling
  - Parallel mmdefragfs, Mmfsck improvements

ESSL and Parallel ESSL

- Parallel ESSL 3.3 for AIX and Linux
  - BLACS Disjoint and Overlapping Process Grids
  - New and Enhanced Eigen value subroutines
    - New: PDSYEVD, PDSYNTRD, PDSYNGST
    - Improved Performance: PDSYEVX, PDSYGVX
- ESSL 4.2.5 for Linux SLES 10
- ESSL 4.3 for AIX and Linux
  - Power6, VMX,
  - Serial and SMP Libraries with 8 byte Integer arguments
  - New LAPACK Linear Equation Subroutines
    - c/zpptri
    - s/d/c/zpocon, s/dlansy. c/zlanhe
    - s/d/c/zppcon. s/dlansp, c/zlanhp
    - s/d/c/zgecon, s/d/c/z/lange
    - sgeqrf, c/zgeqrf

GPFS Information Lifecycle Management

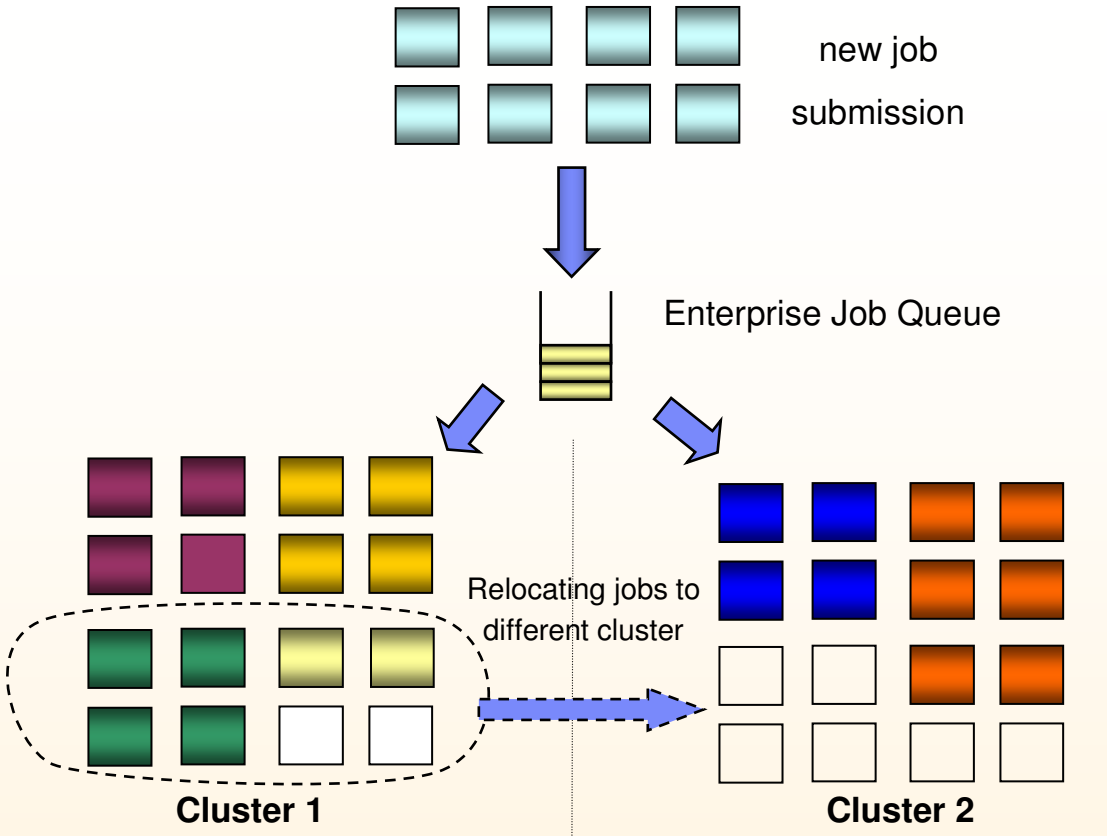
- GPFS supports Information Lifecycle Management (ILM) via three new abstractions: storage pools, filesets, and policies
  - Storage pool – group of LUNs
  - Fileset: subtree of a file system
  - Policy – rule for assigning files to storage pools
- Types of policy rules
  - **Placement**, e.g. place database files on RAID1 storage, place other data files on RAID6
  - **Migration**, e.g. move project files to SATA storage after 30 days, to tape after 60 days
  - **Deletion**, e..g delete scratch files after 7 days



LL Enhancements

- Faster job launch – enhanced pipelining and parallelism
- Dynamic control of SMT on a per job basis
- Modern Web based UI integrated into the ISC framework
- Blue Gene enhancements
  - Fair share scheduling
  - Reservation in advance
- Multi-cluster enablement – Extend domain to the data center
  - Hooks with GPFS ILM in the future

Example demonstrating Enterprise Operational Efficiency and the Scale Across Vision



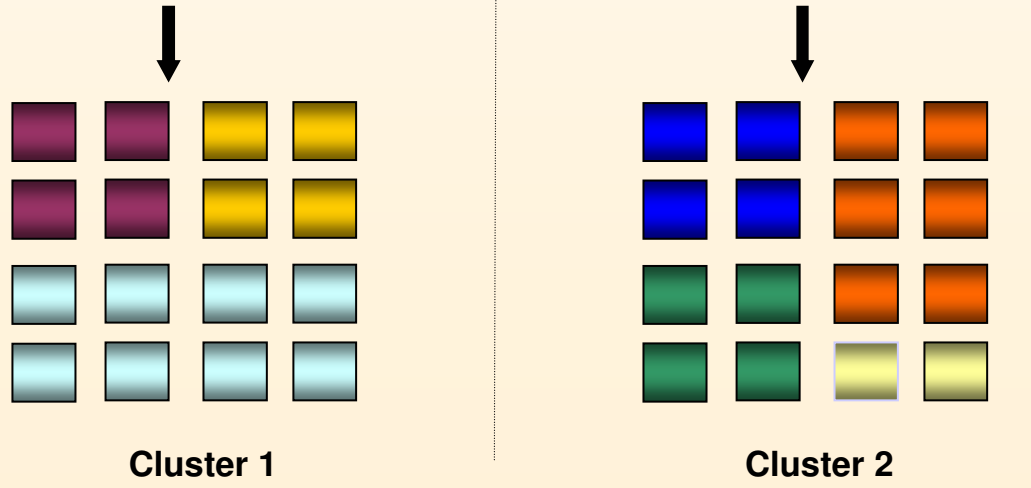
Without C/R and Relocation

Cluster 1: 4 jobs running. 2 nodes free

Cluster 2: 2 jobs running, 6 nodes free

Utilization: Cluster 1: 87%, Cluster 2: 63%

Enterprise Utilization: 75%



With C/R and Relocation

Cluster 1: 3 jobs running. 0 nodes free

Cluster 2: 4 jobs running, 0 nodes free

Utilization: Cluster 1: 100%, Cluster 2: 100%

Enterprise Utilization: 100%